All You Need is Supervised Learning

From Imitation Learning to Meta-RL With Upside Down RL

Upside down RL flips the use of the return in the objective in RL, taking returns as input and predicting actions

Trained using supervised learning: simple for offline, akin to expectation maximisation for online

Commands, c, are any computational predicates that are consistent with the data, for example, the desired time horizon, d^H, and return, d^R

Removing d^R recovers IL, adding g recovers GCRL; but further predicates can be used for self-supervised learning

Recurrent nets solve POMDPs, set-equivariant nets allow dynamic observations/actions/commands - enabling a single model to solve all RL tasks

One Model
LSTM + Perceiver IO

One Loss
Cross-entropy

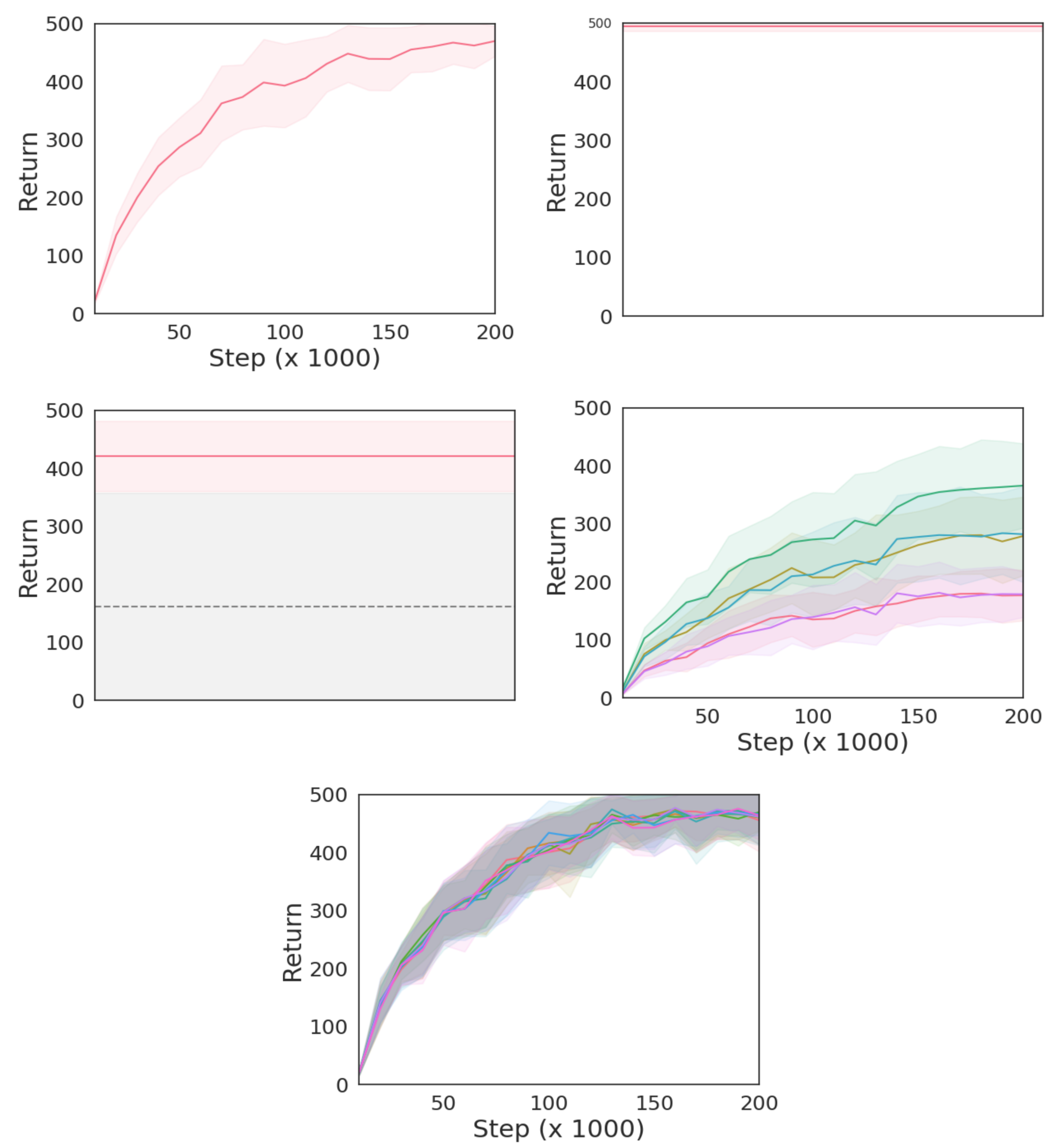
One Algorithm

```
Require: environment E, policy π(a|o,c,h), memory D

function reset(E,π,D)
    Reset environment E and π's hidden state h
    Get initial observation and goal (o,g) from E
    Sample c based on D and (o,g)

Train π on batches from D
if performing IL or offline RL without environment interaction then return

reset(E,π,D)
while true do
    Act with a,h ~ π(a|o,c,h)
    Observe (o',r,g',1_terminal) from environment transition
    Update D with (o,a,r,g,1_terminal)
    Update h (to contain a and r) and c
    Train π on batches from D
    if 1_terminal then reset(E,π,D)
```



Experiments on CartPole variants: online RL, IL, offline RL, GCRL, meta-RL

Minimalist codebase to facilitate further exploration of UDRL!

ARAYA
Imperial College London
USI/SUPSI Istituto Dalle Molle di studi sull'intelligenza artificiale IDSIA
nnaisense

Kai Arulkumaran, Dylan R. Ashley,
Jürgen Schmidhuber & Rupesh K. Srivastava

Paper

Code